Eur J Vasc Endovasc Surg **35**, 295–300 (2008)
doi:10.1016/j.ejvs.2007.09.017, available online at http://www.sciencedirect.com on ScienceDirect

Interobserver and Intraobserver Variability of Interpretation of CT-angiography in Patients with a Suspected Abdominal Aortic Aneurysm Rupture

L.L. Hoornweg,[1] W. Wisselink,[2] A.C. Vahl,[3] J.A. Reekers,[4] O.M. van Delden,[4] D.A. Legemate[1] and R. Balm[1]*

[1]Departments of Vascular Surgery, and [4]Radiology, Academic Medical Center, Amsterdam, The Netherlands,
[2]Department of Vascular Surgery, Vrije Universiteit Medical Center, Amsterdam, The Netherlands,
[3]Department of Vascular Surgery, Onze Lieve Vrouwe Gasthuis, Amsterdam, The Netherlands

Purpose. *To assess interobserver and intraobserver agreement on presence of rupture and determining suitability for endovascular repair (EVAR) on CT angiography (CTA) of patients with a suspected ruptured abdominal aortic aneurysm (RAAA).*
Methods. *For the Amsterdam Acute Aneurysm study, a randomised multicenter trial (ISRCTN66212637), we register all patients with suspected RAAA in the Amsterdam region. For the current analysis 51 consecutive patients were included from this prospective database. Pre operative CT scans were assessed twice with a six-week interval by three vascular surgeons and two interventional radiologists. Variables recorded were presence of rupture, diameter and length of the infrarenal aortic neck, diameters of both iliac arteries and final judgement on anatomical suitability for EVAR. Kappa values for dichotomous outcomes were calculated as a measure of agreement above chance. Continuous outcomes were investigated by calculating the intraclass correlation coefficient (ICC) and by Bland-Altman plots.*
Results. *For presence of rupture group kappa was 0.59 (CI: 0.42–0.77). The group kappa of suitability for EVAR was 0.38 (CI: 0.24–0.51). The ICC for diameter and length of the infrarenal aortic neck and diameters of left and right iliac arteries were 0.40 (CI: 0.26–0.56), 0.47 (CI: 0.32–0.62), 0.61 (CI: 0.48–0.74) and 0.35 (CI: 0.21–0.50) respectively. The Bland-Altman plots confirmed the large variation among observers.*
Intraobserver kappa ranged from 0.57–0.78 for presence of rupture and 0.40–0.80 for suitability for EVAR.
Conclusion. *Moderate interobserver agreement was found for presence of rupture and fair agreement for suitability for EVAR. Intraobserver agreement ranged from moderate to almost perfect. Based on this data, optimization of the protocol is mandatory to identify uniform measuring techniques of well defined anatomical criteria for endovascular repair of ruptured aneurysms.*

Keywords: Ruptured abdominal aortic aneurysm; Endovascular repair; Variability; CTA measurements.

Introduction

Endovascular repair (EVAR) is an emerging treatment for patients with ruptured abdominal aortic aneurysms (RAAA). CT angiography (CTA) plays an important role in selecting patients with suspected RAAA for EVAR, firstly to confirm rupture, secondly to assess anatomical suitability. Anatomical suitability depends on the type of stent-graft used and morphology of the affected abdominal aorta and iliac arteries. Studies using similar stent-grafts and anatomical criteria report different suitability rates.[1–3] However, studies to assess agreement between experienced clinicians on determining presence of rupture and suitability for EVAR in patients with RAAA are lacking.

We studied interobserver and intraobserver variability on determining presence of rupture, diameter and length measurements of the infra renal aortic neck and iliac landing zones and final suitability for EVAR on CTA's of patients with suspected RAAA.

*Corresponding author. R. Balm, MD, PhD, Department of Vascular Surgery G4-107, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, P.O. Box 22700, 1100 DE Amsterdam, the Netherlands.
E-mail address: r.balm@amc.nl

1078–5884/000295 + 06 $34.00/0

Patients and Methods

The Amsterdam Acute Aneurysm trial is a multicenter randomized trial comparing open with endovascular treatment in patients with RAAA (ISRCTN66212637).[4] Background, design and methods of this trial have previously been described.[5] As part of the Amsterdam Acute Aneurysm trial, we register and evaluate all patients with a suspected RAAA in the Amsterdam region (1.3 million inhabitants). In this trial all patients with suspected RAAA are transported to one of three trial centers A, B or C. All hemodynamically stable patients presented at the emergency room and in whom the diagnosis of rupture was not rejected immediately, underwent CT scanning (Protocols are listed in the appendix). From April 8th 2004 till august 29th 2005 we included 62 consecutive patients with suspicion of RAAA and in whom CTA was performed. Scans were not equally divided among the three trial centers. In order to equalize the number of scans per center and to prevent selection bias we decided to study the first 17 consecutive CTA's from center A, the first 16 consecutive CTA's from center B and the first 18 consecutive CTA's from center C (52 CTA's in total).

In order to evaluate interobserver and intraobserver variability we asked 5 observers (3 vascular surgeons and 2 interventional radiologists, all experienced in the endovascular technique) to assess the preoperative CT scans on presence of rupture and anatomical suitability for endovascular repair. We defined RAAA as an aneurysm with presence of intra or retroperitoneal haematoma on CTA.

We asked the observers to judge the presence or absence of rupture. In addition they had to measure diameter and length of the infra renal aortic neck cranial to the aneurysm, the diameter of iliac artery at the location of the proposed landing zone of the endovascular graft and the diameter of the contra lateral common iliac artery. Unequivocal instructions on measuring lengths and diameters were given before the first CT assessment session. Firstly to measure the absolute length of the infrarenal aortic neck that could be used to fix the endovascular graft, secondly to measure the diameter of the infrarenal aortic neck at the proposed landing zone of the endovascular graft. For measuring the iliac artery diameters, the observers could decide in accordance with one's own judgement at what location of the iliac arteries and at which side the endovascular graft should land. Final suitability for the endovascular procedure was determined according to anatomical criteria for a standard set of aorto-uni-iliac endovascular grafts (Talent®, Medtronic, Santa Rosa, CA).[6] The following criteria had to be met: the infrarenal aortic neck cranial to the aneurysm must have a diameter between 20 and 32 mm and a length of at least 10 mm; at least one iliac artery should be able to accommodate the endovascular graft and have a diameter of 8 to 18 mm; the contra lateral common iliac artery diameter should have a diameter between 8 and 20 mm.

Scans (36 digital, 15 hardcopy prints) were made anonymous and axial images of hardcopy prints were scanned and digitized. All scans were presented on a computer with eFilm Lite® (Merge healthcare, Milwaukee, USA), an application used for viewing digitized medical images in DICOM format (Digital Imaging and Communications in Medicine). When viewing images, users can perform adjustments of window width and level, and measure distances on axial images. The quality of the scans was similar to that used in clinical practice. Each observer assessed the scans independently and without time restrictions. All observers were blinded to each other's results and after the first review session no feedback was given. To evaluate reliability of individual assessment of CT angiographic measurements on diagnosing rupture and determining suitability for endovascular repair, the observers reviewed all scans under similar conditions 6 weeks after the first reading session and scans were presented in a different order.

Statistics

Interobserver agreement for dichotomous outcomes was calculated using data from the first review session. Group kappa values were calculated with a dedicated software program: AGREE version 7 (Science plus Group, Groningen, The Netherlands).[7] For intraobserver agreement we used SPSS 12.0 (Microsoft Corporation, USA). The kappa value was interpreted according to the method of Landis and Koch;[8] poor (0 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60), good (0.61 to 0.80) and perfect (0.81 to 1) agreement. No guidelines exist as to which level agreement is acceptable. For continuous data we calculated the intraclass correlation coefficient (ICC, two way mixed, random effect model, absolute agreement) with data from the first review session, using SPSS 12.0. The ICC is the parametric analogue of the chance-corrected kappa measurement of agreement and is also interpreted according the method of Landis and Koch.[8] The ICC not only assesses the strength of linear correlation between two measurements, but also detects systematic error.[9,10] Additionally, Bland Altman plots were made to show the difference between two measurements against their mean, to check whether the error of measurement was independent of the magnitude of the mean score.[11] If measurements could not be made, because of absence of an aneurysm for example, the observers registered “not measurable”. These data were extracted from the Bland Altman plots.

Results

Interobserver agreement

An example of interobserver agreement between observer 1 and 2, and observer 1 and 3 on the presence of rupture is presented in Table 1a and 1b. There were no missing data. Table 2 presents kappa values among all 5 observers on presence of rupture. The overall group kappa value was 0.59 95% CI: 0.42–0.77 for presence of rupture corresponding to moderate agreement.

For each of the anatomical criteria an ICC was calculated. ICC values for diameter and length of the infra renal aortic neck and both iliac arteries are shown in Table 3. For both aortic neck diameter and length agreement was moderate. For left and right iliac artery diameter measurements, agreement was substantial and fair respectively. In Fig. 1 the differences between measurements per observer pair are plotted against the mean of both measurements per observer pair. Fig. 1a visualizes an increase of variability of the measurements as the magnitude of the measurements increases. There is one clear outlier in Fig. 1c. This was caused by the fact that observer 3 chose the external iliac artery as optimal landing zone for the stent-graft, while observer 1 preferred the common iliac artery.

Table 4 represents kappa values amongst all 5 observers on suitability for EVAR. The overall group kappa was 0.38 95% CI: 0.24–0.51, interpreted as fair agreement.

Table 1a. Two-by-two table with ratings of 2 observers on presence of rupture presenting absolute agreement in 43/51 cases (84%), prevalence index of 0.69 and kappa value of 0.43. The boxed values are proportions of agreement

		observer 2		
		+	-	
observer 1	+	39	0	39
	-	8	4	12
		47	4	51

Table 1b. Two-by-two table with ratings of 2 observers on presence of rupture presenting absolute agreement in 46/51 cases (90%), prevalence index of 0.55 and kappa value of 0.72. The boxed values are proportions of agreement

		observer 3		
		+	-	
observer 1	+	37	2	39
	-	3	9	12
		40	11	51

Intraobserver agreement

Tables 2 and 4 show the intraobserver agreement for each of the 5 observers on presence of rupture and suitability for EVAR respectively. The average intraobserver kappa varies from 0.69 to 0.64 indicating a substantial agreement according to Altman.

Discussion

The interobserver agreement was moderate for presence of rupture and fair for suitability for EVAR. The average intraobserver agreement proved to be substantial for both criteria.

This study shows that the interpretation of the CT scan for presence or absence of rupture is not consistent. It should be noted that we did not inform the observers about the patients’ actual medical history and results of physical examination. Clinical information provides pre-CTA likelihood in judging the presence or absence of rupture. The lack of such information might have effected the interpretation of CTA and could explain the moderate agreement. The observers could only choose between presence and absence of rupture, suspicion of rupture was not an option. In

Table 2. Kappa values among 5 observers on presence of rupture. The bold numbers present the Intraobserver agreement. Average inter and intra observer kappa are expressed at the bottom of the table

Presence of rupture	Kappa value				
	observer				
	1	2	3	4	5
observer 1	**0.78**	0.43	0.72	0.53	0.60
observer 2		**0.57**	0.47	0.40	0.32
observer 3			**0.73**	0.57	0.52
observer 4				**0.68**	0.55
observer 5					**0.67**

Average inter observer kappa is 0.59 (95% CI; 0.42–0.77).
Average intra observer kappa is 0.69 (95% CI; 0.59–0.78).

Table 3. Interobserver agreement of continuous variables between the 5 observers, expressed by intraclass correlation coefficient (ICC)

Criteria	ICC	95% CI
infrarenal neck		
diameter (DN)	0.40	0.26−0.56
length (LN)	0.47	0.32−0.62
iliac artery left		
diameter (DIL)	0.61	0.48−0.74
iliac artery right		
diameter (DIR)	0.35	0.21−0.50

clinical practice 'suspicion' can be weighted with other clinical criteria, and eliminating this option could have decreased agreement. Since not all patients underwent open repair in this study, a reference standard on the presence of rupture is missing. We found one rather dated article in which RAAA was prospectively diagnosed on CT in 17 patients for whom surgical correlation was available.[12] Only 10 out of 17 patients demonstrated an actual rupture during surgery. Unfortunately no data on agreement were reported. The identification of rupture is an important clinical problem involving life-saving decisions. The results of our study show that improved criteria for a consistent diagnosis of RAAA are mandatory.

We also found variability in the decision on suitability for EVAR. Agreement on all anatomical criteria was fair to moderate, except for the diameter of the left iliac artery which was substantial. We are unable to identify one specific anatomic criterion to explain the variability on suitability for EVAR.

Some understanding of the disagreement on suitability for EVAR can be explained by visual interpretation of the Bland Altman plots. Fig. 1a shows that an average measurement of the infra renal aortic diameter of 30 mm on the x-axis, correlates with differences of 2 mm to 5 mm in measurements between the two observers on the y-axis. Since a maximum neck diameter of 32 mm is the cut-off criterion in assessing suitability for endovascular repair, a difference in

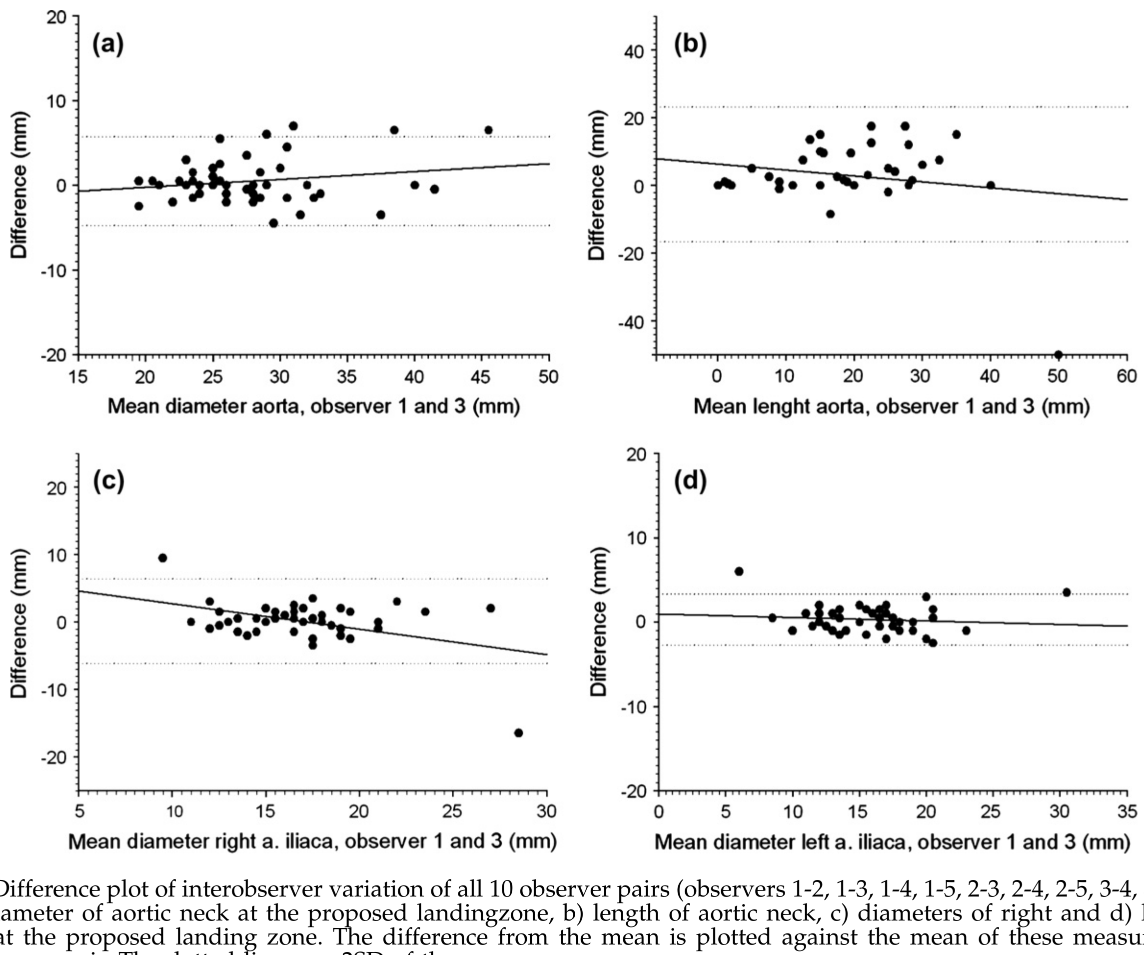

Fig. 1. Difference plot of interobserver variation of all 10 observer pairs (observers 1-2, 1-3, 1-4, 1-5, 2-3, 2-4, 2-5, 3-4, 3-5, 4-5) of a) diameter of aortic neck at the proposed landingzone, b) length of aortic neck, c) diameters of right and d) left iliac artery at the proposed landing zone. The difference from the mean is plotted against the mean of these measurements per observer pair. The dotted lines are 2SD of the mean.

Table 4. Kappa values among 5 observers on suitability for EVAR. The bold numbers present the Intraobserver agreement. Average inter and intra observer kappa are expressed at the bottom of the table

Suitability for EVAR	Kappa value				
	observer				
	1	2	3	4	5
observer 1	**0.72**	0.30	0.51	0.26	0.37
observer 2		**0.40**	0.40	0.22	0.34
observer 3			**0.76**	0.57	0.37
observer 4				**0.80**	0.22
observer 5					**0.54**

Average inter observer kappa is 0.38 (95% CI; 0.24−0.51).
Average intra observer kappa is 0.64 (95% CI; 0.43−0.85).

measurement of 5 mm makes the difference to decide for open or endovascular repair.

A hypothesis for the differences in measurements could be the presentation of the digital and hardcopy scans with eFilm Lite®. Only axial images were presented and the observers could not perform multi planar reconstructions (MPR) or create central lumen lines (CLL), which is possible in the actual clinical setting. Presentation of all scans on a workstation, with the ability to evaluate three dimensional reconstructions, MPR and CLL, might be closer to the actual clinical situation and result in more accurate measurements.

We have to note that neck angle, tortuosity and assessment of calcification were subjective in this study. A clear definition of these variables and recording of this data could have provided more specific information on unsuitability for EVAR.

Rödel *et al.*[13] studied agreement on suitability for elective EVAR between 5 clinicians all experienced with the endovascular technique. Interobserver agreement ranged from kappa value 0.08 to 0.62, underlining the considerable variation between the clinicians on suitability for endovascular repair. In contrast to this study, anatomical data were not measured, but already presented on the worksheet.

In this study kappa values are used to express intra and interobserver agreement. Kappa has some limitations and is sensitive to an extreme distribution of the two-by-two table. If proportions of agreement on positive classifications differ from proportions of agreement on negative classification, the kappa value is lower. Table 1a and 1b illustrate this problem. Although absolute agreement only differs 6%, kappa values vary from 0.43 to 0.72 as a result of a more asymmetrical distribution on the proportions of agreement in Table 1a.[14]

Furthermore, kappa is influenced by the prevalence of the presence of rupture. For example, if the prevalence is high, agreement beyond chance is high in advance. In this case the possible agreement beyond chance becomes low with only moderate kappa values. These limitations oblige us to interpret kappa values with care. All observers knew that all CT scans were for patients with suspected RAAA. This will have added considerably to the high prevalence and asymmetry of the 2×2 tables. A possible solution might have been to include a larger proportion of CT scans for known non-ruptured AAA. Theoretically, sensitivity and specificity of the test is more informative, but the absence of a reference standard excludes this possibility.

As ICC is mathematically equivalent to the kappa statistic,[9] limitations as described for kappa values also apply for ICC values and results should be interpreted likewise.

This study shows that the interpretation of the CT scan for presence of rupture and suitability for endovascular repair of RAAA, even if anatomical criteria are well defined, is not consistent. Apart from limitations in interpreting the kappa value and presenting scans with eFilm Lite®, the results of our study are of significant clinical importance and need further research.

Conclusions

Interobserver agreement for presence of rupture was moderate and agreement on suitability for EVAR was fair. Intraobserver agreement ranged from moderate to almost perfect agreement. These results incite to optimization of the protocol and oblige us to identify uniform measuring techniques of well defined anatomical criteria for endovascular repair of ruptured aneurysms.

Scan protocols

Center A
Siemens® Somatom Sensation 4, Erlangen, Germany

slice thickness 5 mm, collimation 10 mm, reconstruction increment 2.5 mm, pitch 1.5

Contrast protocol: power injection 100 ml Visipaque™ (GE Healthcare Limited, Buckinghamshire, UK) 320, 3 ml/s, bolus tracking

Center B
Siemens® Somatom Sensation 64, Erlangen, Germany

slice thickness 1.5 mm, collimation 6 mm, reconstruction increment 1 mm, pitch 1.2

Contrast protocol: power injection 100 ml Ultravist™ (Bayer Healthcare Pharmaceuticals, Germany) 300, 5 ml/s, bolus tracking

Center 3
Philips® Tomoscan AVE 1, Philips Medical Systems, Eindhoven, The Netherlands

slice thickness 7 mm, collimation 7 mm, reconstruction increment 5 mm, pitch 1.4

Contrast protocol: power injection 100 ml Omnipaque™ (GE Healthcare Limited, Buckinghamshire, UK) 300, 2 ml/s, delay 50 sec.

Acknowledgements

The authors would like to thank Jan Wolters for preparing the scans for review.

Funding

This work is being funded by a research grant (grant 2002B197) from the Dutch Heart Foundation, The Hague, the Netherlands.

References

1 Hoornweg LL, Wisselink W, Vahl AC, Balm R. Amsterdam Acute Aneurysm Trial Collaborators. The Amsterdam Acute Aneurysm Trial: Suitability and Application Rate for Endovascular Repair of Ruptured Abdominal Aortic Aneurysms. *Eur J Vasc Endovasc Surg* 2007;**33**:679–683.

2 Peppelenbosch N, Geelkerken RH, Soong C, Cao P, Steinmetz OK, Teijink JA *et al*. Endograft treatment of ruptured abdominal aortic aneurysms using the Talent aortouniiliac system: an international multicenter study. *J Vasc Surg* 2006;**43**(6):1111–1123.

3 Reichart M, Geelkerken RH, Huisman AB, van Det RJ, de Smit P, Volker EP. Ruptured abdominal aortic aneurysm: endovascular repair is feasible in 40% of patients. *Eur J Vasc Endovasc Surg* 2003;**26**(5):479–486.

4 http://www.controlled-trials.com/isrctn/trial/|/0/66212637.html. 12-9-2005. Ref Type: internet Communication.

5 Amsterdam Acute Aneurysm Trial Collaborators. Amsterdam Acute Aneurysm Trial; Background Design and Methods. *Vascular* 2006;**14**(3):1–6.

6 Peppelenbosch N, Zannetti S, Barbieri B, Buth J. Endograft treatment in ruptured abdominal aortic aneurysms using the Talent AUI stentgraft system. Design of a feasibility study. *Eur J Vasc Endovasc Surg* 2004;**27**(4):366–371.

7 www.scienceplus.nl. 2007. Ref Type: internet Communication.

8 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**(1):159–174.

9 Fleiss JL, Cohen RA. The equivalent of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973;**33**:613–619. Ref Type: Abstract.

10 Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis* 1985;**38**(1):27–36.

11 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;**8**(2):135–160.

12 Gale ME, Johnson WC, Gerzof SG, Robbins AH. Problems in CT diagnosis of ruptured abdominal aortic aneurysms. *J Comput Assist Tomogr* 1986;**10**(4):637–641.

13 Rodel SG, Geelkerken RH, van Herwaarden JA, Kunst EE, van den Berg JC, van der PJ *et al*. Consistency in endovascular aneurysm repair suitability assessment requires group decision audit. *J Vasc Surg* 2006;**43**(4):671–676.

14 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005; **85**(3):257–268.

Accepted 25 September 2007
Available online 7 November 2007